Faculty Working Papers

INTERTECHNIQUE CROSS-VALIDATION IN CLUSTER ANALYSIS

A. Marvin Roscoe, Jagdish N. Sheth,
and Welling Howell

#175

College of Commerce and Business Administration
University of Illinois at Urbana-Champaign

FACULTY WORKING PAPERS

College of Commerce and Business Administration

University of Illinois at Urbana-Champaign

April 4, 1974

INTERTECHNIQUE CROSS-VALIDATION IN CLUSTER ANALYSIS

A. Marvin Roscoe, Jagdish N. Sheth,
and Welling Howell

#175

Intertechnique Cross-Validation in Cluster Analysis

A. MARVIN ROSCOE,

JAGDISH N. SHETH,

and

WELLING HOWELL *

* A. Marvin Roscoe and Welling Howell are Marketing Supervisors in the Market Research Section of the Marketing Department of the AT&T Company. Jagdish N. Sheth is I.B.A. Distinguished Professor and Research Professor at the University of Illinois, Urbana - Champaign.

Intertechnique Cross-Validation in Cluster Analysis

In view of the fact that in practical marketing research clustering methods are utilized to define homogeneous market segments by empirical research, it is critical to ensure that the derived clusters are, in fact, the true clusters. One procedure of ensuring cluster invariance is replication, but this is not always practical. A second procedure common in psychometrics is one of cross-validating the results by external validation.

The objective of this paper is to describe a cross-validation procedure which utilized intertechnique comparisons of the clustering results. The procedure is applied to two hierarchical procedures in which the problem involves the determination of geographical heterogeneity of markets for the telephone industry.

Intertechnique Cross-Validation in Cluster Analysis

Cross-validation among techniques seems essential in cluster analysis because most clustering methods tend to be heuristic algorithms instead of analytically optimal solutions. (See Joyce and Channon [4] and Frank and Green [2] for a review of the numerous clustering methods available today). As heuristic algorithms, they have no sampling theory for statistical inferences about the size and the number of clusters. Also, there are no external validation procedures to ensure that the clusters derived from a specific cluster analysis are in reality the true invariant clusters. The potential statistical problem of obtaining artifacts as clusters is further compounded in some procedures which require a priori assumptions about the size and the number of clusters. Although a number of clustering methods perform statistical tests such as the F ratio or Wilks' Lambda based on analysis of variance principles to guard against obtaining random solutions, no procedure exists which will increase the assurance that a nonrandom cluster solution is in fact the true cluster solution.

Because clustering methods are used in marketing research to identify homogeneous market segments for selective marketing efforts, it is critical that the clusters derived from a heuristic algorithm are the true clusters. One procedure to ensure cluster invariance is replication which, however, is not always practical. Another procedure is the common practice in psychometrics of cross-validating the results by external validation.

The objective of this paper is to describe a cross-validation procedure which utilizes intertechnique comparisions of the clustering results. Although the actual study entailed applications of five different clustering techniques, our discussion is limited to two techniques in this paper due to space limitations. A brief description of the large scale research project is provided in which the clustering results were essential to formulating an experimental design for a field experiment.

DESCRIPTION OF THE STUDY

The major research study consisted of a three factorial-64 cell experimentation on survey research methods. The three factors were: first, two different lengths of the questionnaire; second, four different follow-up procedures; and, third, the market heterogeneity of geographical areas of the United States with respect to consumer telephone behavior and socioeconomic-demographic characteristics (see [6]). The levels of the first two factors were predetermined based on theory, prior research and practical implications for the ongoing research on a longitudinal national panel of telephone customers. For the third factor, it was necessary to determine the heterogeneity of the markets by empirical research which utilized clustering methods.

To define the market heterogeneity, profile data on 30,000 residential telephone customers were used for clustering. These customers are part of a longitudinal consumer panel called the Marketing Research Information System which is maintained for the Bell System by AT&T. The panel members are selected based on a multistaged stratified sample in which the first stage of the sampling procedure consists of 100 Revenue Accounting Offices (RAOs) representing the entire Bell System. The profile consists of essentially three types of information about each panel member:

(a) his socioeconomic - demographic status and housing characteristics determined by a survey conducted in early 1970 and matched with the 1970 Census, (b) his monthly telephone behavior broken down into several categories as determined by the industry practice, and (c) an inventory of his telephone equipment including number and types of telephones, and additional services.

Since it was required to empirically investigate the geographical heterogeneity of the markets, an average profile of the residential telephone customers was determined for each of the 86 RAOs for which detailed and complete information was available.

A total of 65 customer descriptors were used to represent the total profile of customers. A list of the variables is shown in Table 1. A factor analysis (principal components) solution with orthogonal Varimax rotation was performed on the data for the following reasons: (a) to reduce the multicollinearity among variables so that the profile consisted of orthogonal factor scores which are geometrically essential to calculate Euclidian distances, (b) to equalize the relative weights of each of the underlying dimensions which could otherwise be easily changed by arbitrary dropping or adding of profile variables, and (c) to standardize the diverse scales of measurement common across the socioeconomic, demographic and telephone information [5]. Ten significant factors were extracted from the analysis which summarized 92 percent of the total variance. A brief description of the factors is provided in Table 2.

The number of significant factors was determined using several criteria, both statistical and judgmental, following the recommendations of Rummel [7]. In addition, the stability of the factor structure was also determined by comparing the results with other data analyses to ensure the invariance of the fundamental dimensionality and structure of the profile data.

The standarized rotated factor scores for each RAO were then utilized to compute Euclidian distances between all combinations of RAOs. The resultant 86 X 86 distance matrix became the input to the clustering procedures.

Due to the following distinct advantages, Johnson's Hierarchical Clustering method [3] was chosen as the primary clustering technique for determining the market heterogeneity. First, it is strictly empirical; second, no prior assumptions are required on the part of the researcher; and third, a hierarchical display is provided of the clusters being formed based on a function minimizing the pairwise distances among entities. While the size of the distance matrix is a limitation of the technique, it was not a problem in our case because of the relatively small number of RAOs to be clustered. Due to the structure of the distance matrix and the presumption of the "ultrametric inequality", [3, p. 248-9] the diameter method was chosen instead of the connectedness method in the BE-HICLUST solutions. The results are diagramed in Figure 1.

While the hierarchical clusters from HICLUST were meaningful and had strong face validity, it was necessary to cross-validate the results by at least one other technique which was essentially similar in its input requirements, analytic strategies and the output format. For this we chose the cluster analysis program developed as part of the BMDP Series which is also a hierarchical clustering routine based on sum of squares distances and the amalgamation principle [1]. In short, BMDP2M amalgamates entities based on the criterion of the smallest distance. Once a cluster is formed, consisting of at least two entities, it calculates the average profile of the cluster and treats it as if it were a new entity which is then clustered with other entities or clusters based on the principle of smallest distances. The process continues until all entities and clusters are hierarchically linked at different levels of distances. The results of the BMDP2M analysis are diagramed in Figure 2.

As can be seen, the two hierarchical clusters are similar in their structure and hierarchy suggesting that there is a good cross-validation between the two analyses. In order to quantitatively assess the degree of congruence between the two hierarchical clusters, two distinct statistical procedures were utilized. The first procedure consisted of calculating the correlation coefficient for the two distributions of distances at which linkages were made between entities or clusters in each hierarchical analysis. Since the number of linkages is not likely to be identical, we have selected the maximum number of links of one technique and the corresponding number of the other technique. The correlation coefficient between the sequential linkage distances is 0.994. which is highly positive indicating extreme closeness of the hierarchical structure of the two cluster analyses.

Another procedure for cross-validation consisted of examining the clusters developed at some specific levels of distances. Based on the plotting of distances at which linkages were made, for the BE-HICLUST results a distance of 5.00 was indicated as a cutoff point due to the natural break in the curve suggesting a clear truncation.

The linkage for the BMDP2M results were also plotted and the natural break in the linkages occurred at 3.1. This was at the point where all the clusters had been formed. After this point the BMDP2M analysis indicated 15 unique entities that were not identified with any of the defined clusters. In order to produce comparable results, the cutoff point for the BE-HICLUST diagram was moved to 3.5 for the cross-validation. The clusters could be identified by their geographical orientation and have been labeled Eastern, Southern, Central and Western. Metropolitan has been used for large urban areas not specifically associated with regional areas. The clusters derived from the two techniques are marked in Figures 1 and 2 and are cross-tabulated in Table 3.

A total of 17 clusters are displayed in Table 3, consisting of 13 regional clusters (Eastern, Southern, Central and Western), three metropolitan cities clusters and the last one representing all the unique RAOs which could not be clustered due to their extreme distances from other RAOs. The cross-tabulation between HICLUST and BMDP2M clustering results indicates that 62 out of 86 RAOs fell on the diagonal of the crosstab matrix which represents a hit of 72 percent correct classifications in terms of intertechnique results. Futhermore, most of the off-diagonal elements generally fall across clusters within the same geographical region. In Table 4, a cross-tabulation at the regional level is provided which shows that 75 out of 86 RAOs could be correctly classified on an intertechnique basis. This represents a hit of 72 percent.

While the two results are quite comparable, there are differences in the example worth noting. The BE-HICLUST algorithm appears to provide a more logical structure to the clusters which are grouped by region as indicated in Figure 2. In addition, the BE-HICLUST method seems to work better where large distances are involved, associating 8 of the 14 unique entities with meaningful clusters. Such differences reinforce the need to use several techniques and to understand the advantages of each especially where the researcher's judgement plays such an important role.

SUMMARY AND CONCLUSIONS

We have pointed out the need for intertechnique cross-validation in cluster analysis due to the heuristic nature of most clustering procedures and the judgemental decisions required to interpret the results. In this paper, we have also presented a concrete application of two statistical procedures which enable the researcher to quantitatively measure the congruence of structure and content of clusters across techniques. The first consists of a correlation coefficient index calculated on the distributions of distances at which sequential linkages are made among entities or clusters or both. The second consists of a cross-tabulation of specific clusters derived across two different solutions. In this paper the intertechnique cross-validation procedures have been applied with respect to two hierarchical clustering procedures in which the problem was the determination of geographical heterogeneity of markets for the telephone industry.

Table 1

LIST OF VARIABLES

Housing

1. Own-rent home
2. Type of residence
3. Number of rooms

Mobility

4. Length of residence

Head of Household

5. Sex
6. Age
7. Education
8. Occupation

Family

9. Income
10. Number in family
11. Average Age
12. Life cycle
13. SES status

Telephone Service and Equipment

14. Class of service
15. Grade of service
16. Number of telephones
17. Number of vertical services

Billing Items 12 months

18-29 Local service
30-41 Local message
42-53 Intrastate long distance
54-65 Interstate long distance

Table 2

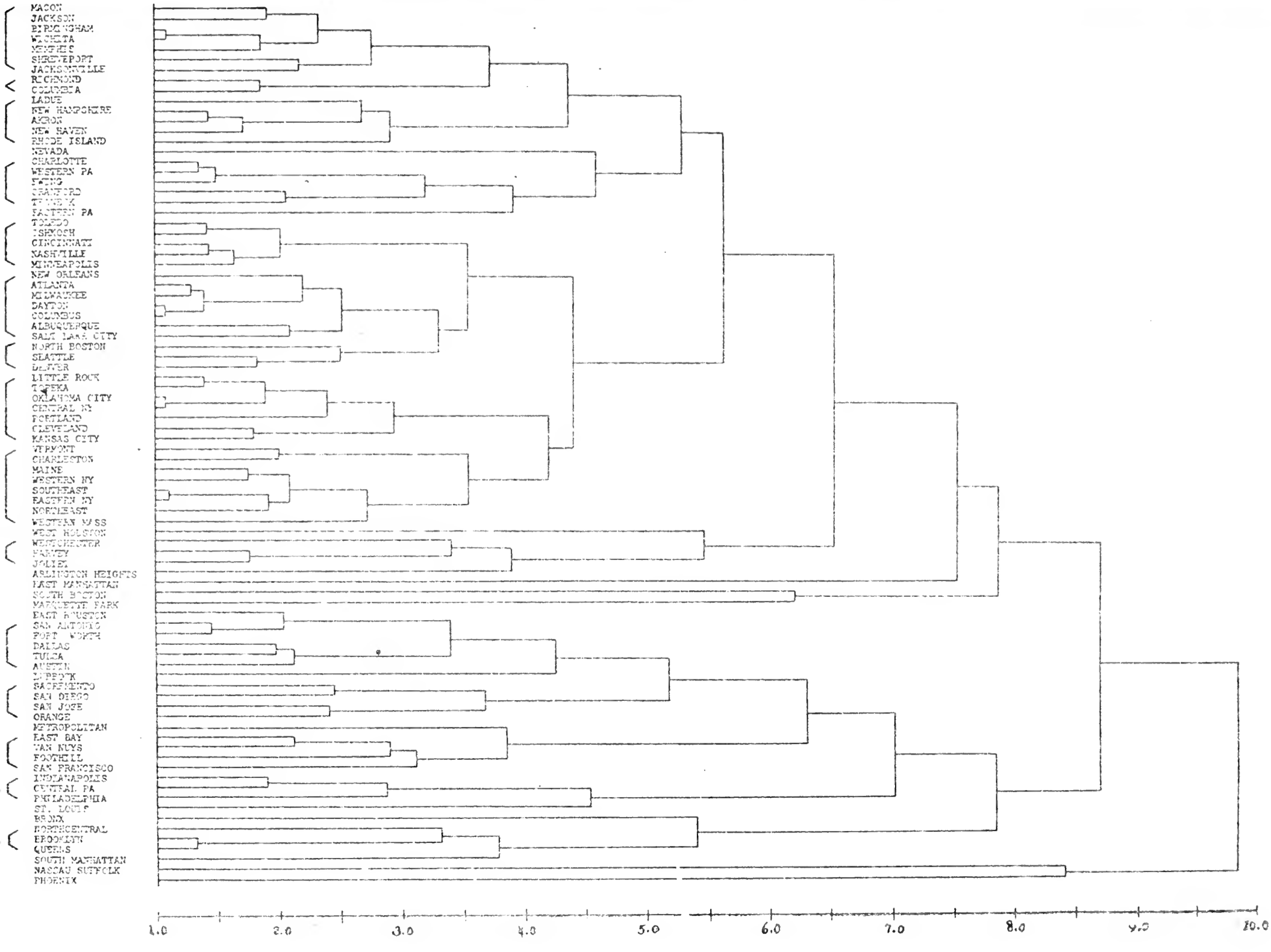
FACTOR DIMENSION LABELS

1. Local service billing
2. Local message billing
3. Intrastate long distance
4. Family - housing
5. Intrastate long distance
6. Life cycle
7. Service and equipment
8. Interstate Long Distance 1 *
9. Interstate Long Distance 2 *
10. Socioeconomic Characteristics

* The two factors for interstate long distance represent different seasonal patterns of calling across geographical areas.

Figure 1

HIERARCHICAL STRUCTURE OF TELEPHONE AREAS - BE-HICLUST



HIERARCHICAL STRUCTURE OF TELEPHONE AREAS - BMDP2M

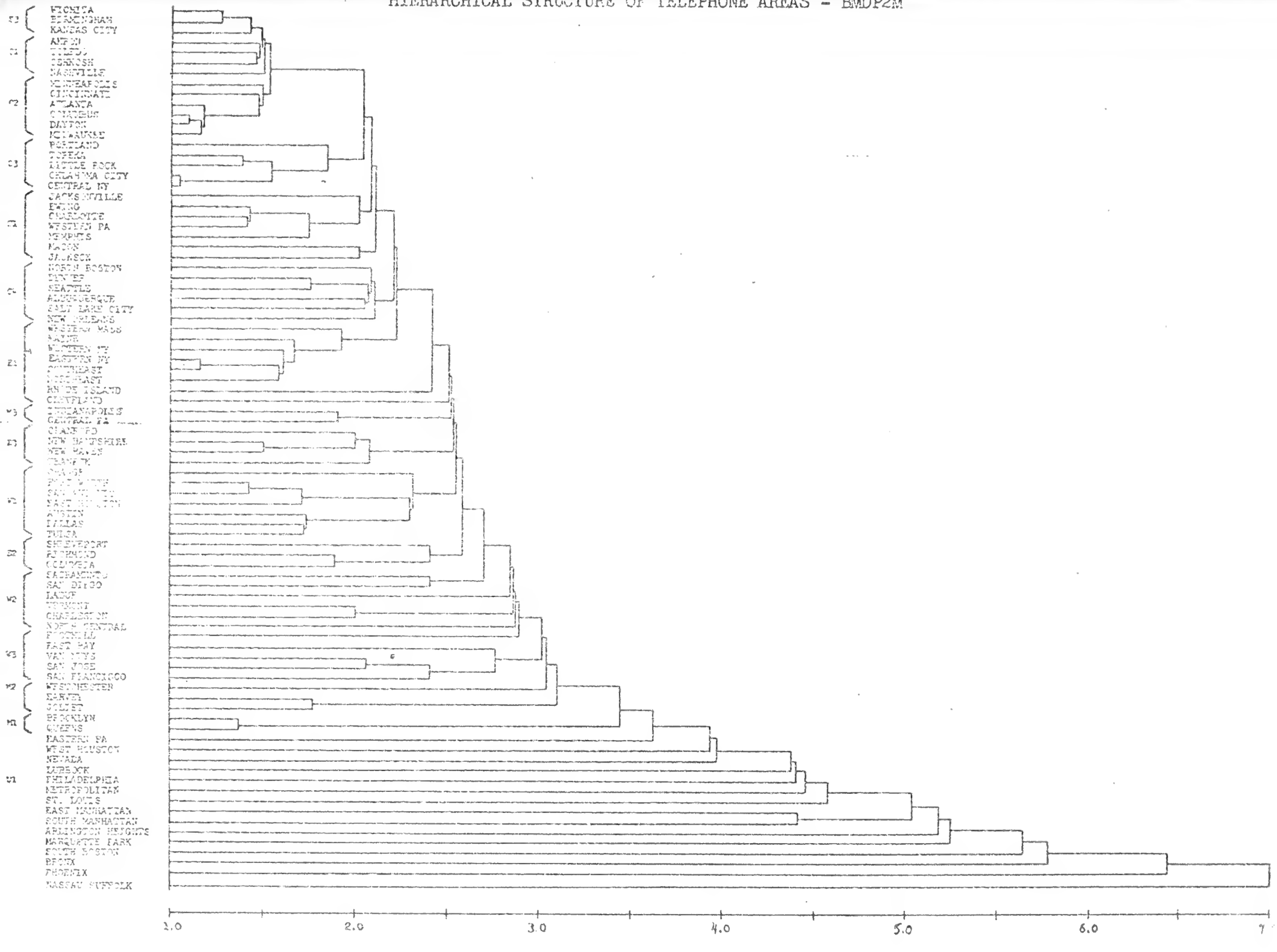


Table 3

CROSS-TABULATION OF CLUSTERS

RE-HICLUST BMDP2M

		EASTERN 1	EASTERN 2	EASTERN 3	SOUTHERN 1	SOUTHERN 2	SOUTHERN 3	CENTRAL 1	CENTRAL 2	CENTRAL 3	CENTRAL 4	WESTERN 1	WESTERN 2	WESTERN 3	METROPOLITAN 1	METROPOLITAN 2	METROPOLITAN 3	UNIQUE RAOS
EASTERN	1	6											2					
	2	1	0	2				1					1					
	3			2	3													
SOUTHERN	1				4	1	2											
	2																	
	3																	
CENTRAL	1							3	2									
	2								4		3							
	3	1					1			5								
	4										3							
WESTERN	1											6						
	2											1	2	1				
	3													4				
METROPOLITAN	1												1		2			
	2															3		
	3																2	1
UNIQUE RAOS																		14

Table 4

CROSS-TABULATION OF REGIONAL, METROPOLITAN AND UNIQUE CLUSTERS

BE-HICLUST BMDP2M

	EASTERN	SOUTHERN	CENTRAL	WESTERN	METROPOLITAN 1	METROPOLITAN 2	METROPOLITAN 3	UNIQUE RAOS
EASTERN	11	3	1	3				
SOUTHERN		9						
CENTRAL	1	1	20					
WESTERN				14				
METROPOLITAN 1				1	2			
METROPOLITAN 2						3		
METROPOLITAN 3							2	1
UNIQUE RAOS								14

REFERENCES

1. Dixon, W.J. "BMD P Series Documentation," Health Sciences Computing Facility. Los Angeles: University of California.
2. Frank, Ronald B. and Paul E. Green. "Numerical Taxonomy in Marketing Analysis: A Review Article," Journal of Marketing Research, 5 (February 1968), 83-98.
3. Johnson, Stephen C. "Hierarchical Clustering Schemes," Psychometrika, 32 (September 1967), 241-54.
4. Joyce, Timothy and C. Channon. "Classifying Market Segment Respondents," Applied Statistics, 15 (November 1966), 191-215.
5. Morrison, Donald G. "Measurement Problems in Cluster Analysis," Management Science, 13 (August 1967), B775-80
6. Roscoe, A. Marvin, Dorothy Lang and Jagdish N. Sheth. "Experimental Effects of Follow-up Methods, Questionnaire Length, and Market Heterogeniety in Mail Surveys", Manuscript submitted for publication.
7. Rummel, R.J. Applied Factor Analysis. Evanston: Northwestern University Press, 1970, Chapter 15.